# Prediction of Lymph Node Maximum Standardized Uptake Value in Patients With Cancer Using a 3D Convolutional Neural Network: A Proof-of-Concept Study

Hiram Shaish[1]
Simukayi Mutasa[1]
Jasnit Makkar[1]
Peter Chang[2]
Lawrence Schwartz[1]
Firas Ahmed[1]

**OBJECTIVE.** The purpose of this study is to determine whether a convolutional neural network (CNN) can predict the maximum standardized uptake value ($SUV_{max}$) of lymph nodes in patients with cancer using the unenhanced CT images from a PET/CT examination, thus providing a proof of concept for potentially using deep learning to diagnose nodal involvement.

**MATERIALS AND METHODS.** Consecutive initial staging PET/CT scans obtained in 2017 for patients with pathologically proven malignancy were collected. Two blinded radiologists selected one to 10 lymph nodes from the unenhanced CT portion of each PET/CT examination. The $SUV_{max}$ of the lymph nodes was recorded. Lymph nodes were cropped and used with the primary tumor histology type as input for a novel 3D CNN with predicted $SUV_{max}$ as the output. The CNN was trained using one cohort and tested using a separate cohort. An $SUV_{max}$ of 2.5 or greater was defined as FDG avid. Two blinded radiologists separately classified lymph nodes as FDG avid or not FDG avid on the basis of unenhanced CT images and separately using a short-axis measurement cutoff of 1 cm. Logistic regression analysis was performed.

**RESULTS.** A total of 400 lymph nodes (median $SUV_{max}$, 6.8 [interquartile range {IQR}, 2.7–11.6]; median short-axis, 1.1 cm [IQR, 0.9–1.6 cm]) in 136 patients were used for training. A total of 164 lymph nodes (median $SUV_{max}$, 3.5 [IQR, 1.9–8.6]; median short-axis, 1.0 cm [IQR, 0.7–1.4 cm]) in 49 patients were used for testing. The predicted $SUV_{max}$ was associated with the real $SUV_{max}$ (β estimate = 0.83, $p < 0.0001$). The predicted $SUV_{max}$ was associated with FDG avidity ($p < 0.0001$), with an ROC AUC value of 0.85, and it improved when combined with radiologist qualitative assessment and short-axis criteria.

**CONCLUSION.** A CNN is able to predict with moderate accuracy the $SUV_{max}$ of lymph nodes, as determined from the unenhanced CT images and tumor histology subtype for patients with cancer.

**Keywords:** convolutional neural network, lymph nodes, machine learning, oncology, PET/CT

doi.org/10.2214/AJR.18.20094

Received May 2, 2018; accepted after revision July 18, 2018.

[1]Department of Radiology, Columbia University Medical Center, 630 W 168th St, New York, NY 10016. Address correspondence to H. Shaish (hs2926@cumc.columbia.edu).

[2]Department of Radiology and Biomedical Imagery, University of California San Francisco, San Francisco, CA.

**Supplemental Data**
Available online at www.ajronline.org.

*AJR* 2019; 212:238–244

0361–803X/19/2122–238

Lymph node metastases occur in many primary malignancies, and their presence portends a worse prognosis [1–3]. Imaging modalities including CT, MRI, and FDG PET/CT have all been used for initial staging of malignancies arising in the chest, abdomen, and pelvis. Each modality has varied and imperfect sensitivities and specificities for evaluating nodal involvement and distinguishing between malignant and benign lymph nodes [4–13]. Detection of nodal metastases is of paramount importance for most treatment planning [1, 2]. Many criteria and imaging analysis tools have been developed for classifying lymph nodes as benign or malignant. These include short-axis measurement [3, 14–16], shape [3, 17], texture analysis [17, 18], CT attenuation [19–21], and standardized uptake value (SUV) cutoff [22–26]. The methods used for texture analysis include machine learning using classic methods and, more recently, deep learning through the use of convolutional neural networks (CNNs) [18, 27, 28]. The purpose of the present study is not to replace PET by predicting SUVs from CT images but, rather, to begin to evaluate whether information available from CT images alone potentially could be informative and complementary to information available from PET.

To date, studies have focused on classifying lymph nodes as benign or malignant, with histopathologic analysis used as the reference standard. The major challenges associated with this approach are the need for accurate radiologic-pathologic correlation and the inherent selection bias of obtaining the pathologic diagnosis from only a few select lymph nodes. As a result, many of the lymph nodes visualized in a patient's body on cross-sectional imaging have not, to our knowledge, been previously studied in this context. In addition, lymph nodes may range from being affected by micrometastatic disease to being completely replaced by tumor. Therefore, analysis of lymph node metastases as a continuous variable correlating with increasing tumor volume may be a more accurate representation than the traditional binary benign-versus-malignant classification.

The maximum SUV ($SUV_{max}$) and the metabolic tumor volume from PET/CT have been shown to correlate with pathologic tumor volume [29]. Unlike lymph node–specific histopathology, the $SUV_{max}$ is readily available for all visible lymph nodes in patients undergoing PET/CT. Although many factors contribute to the $SUV_{max}$ of a given lymph node, the accessibility and established role of the $SUV_{max}$ in clinical practice as well as its quantitative nature make it an easily studied surrogate of malignancy.

We hypothesize that information embedded within the unenhanced CT images of individual lymph nodes, which is not always interpretable by humans, may be harnessed through machine learning and CNNs to predict malignancy. Because of the lack of robust, unbiased pathologic data for multiple lymph nodes (both benign and malignant) in these patients, we chose to study the $SUV_{max}$ as a surrogate endpoint for malignancy.

## Materials and Methods

This retrospective HIPAA-compliant study was approved by the institutional review board at Columbia University Medical Center. A waiver of informed consent was obtained.

### Outline

To test our hypothesis, we built and tested a CNN capable of predicting $SUV_{max}$ in lymph nodes found in patients with cancer, with use of data from the unenhanced CT portion of the initial staging PET/CT examination, and we created a regression analysis model that incorporated the CNN output along with short-axis measurement criteria and radiologist qualitative assessment. Our study consisted of a training phase and a testing phase. In the training phase, two CNNs were constructed. The first CNN accepted only the unenhanced CT image data for each lymph node, whereas the second CNN accepted the histopathologic type of the primary tumor as well as the unenhanced CT image data. In the subsequent testing phase, the CNN was tested on an independent cohort of patients and lymph nodes.

The initial dataset was split into training, validation, and testing sets. The training set is the set of images used by the network for initial training. The validation set is a separate set of images used to optimize the neural network parameters. After training the network with the training set and fine-tuning the network parameters on the basis of the network's performance with the validation set, a separate run was performed that used the entirety of the training and validation data as training data. The trained CNN was subsequently evaluated using the testing set, which was sequestered from the rest of the data at the outset of the study and represented a set of examples never seen by the network. The test data are used to provide an unbiased evaluation of a final model generalization performance and avoid the bias introduced by overfitting in cross-validation methods.

A flowchart of the study is shown in Figure 1.

### Training Population

A total of 2486 PET/CT examinations performed for oncologic evaluation during a 10-month period at Columbia University Medical Center were collected by two board-certified radiologists. The electronic medical record of each patient was evaluated for the pathologic diagnosis of the primary tumor and for any history of treatment before PET/CT. Only the studies performed for initial staging were included.

The same two radiologists examined the unenhanced CT images for each included PET/CT examination and selected from one to 10 discrete lymph nodes in consensus, with up to five lymph nodes selected above the diaphragm and five selected below the diaphragm. Any lymph node, regardless of shape or size, was selected. The only reason a visualized lymph node could be excluded was if it was obscured by adjacent soft-tissue structures (organs, muscles, and other lymph nodes). The radiologists remained blinded to the PET images and prospective radiology reports at the time of lymph node selection. This resulted in the selection of 400 lymph nodes (mean, 2.7 lymph nodes per patient; range, one to 10 lymph nodes per patient) in 136 patients (mean age, 64 years; range, 20–93 years).

### Testing Population

A total of 833 different PET/CT examinations performed for oncologic evaluation during a subsequent 3-month period at our institution were collected by the same two radiologists. The same previously described process used for the training population was repeated in terms of PET/CT and patient selection as well as individual lymph node selection. This repeated process resulted in the selection of 164 lymph nodes (mean, 3.3 lymph nodes per patient; range, one to six lymph nodes per patient) in 49 patients (mean age, 67 years; range, 33–89 years). None of the patients included in the testing population were part of the training population.

### Image Acquisition

PET/CT examinations were performed using a Biograph 64 or Biograph 128 scanner (both from Siemens Healthcare). Patients fasted for 6 hours before imaging was performed, and blood glucose levels were checked. Patients with a blood glucose level greater than 200 mg/dL did not undergo scanning. Oral contrast medium was administered to all patients. Images including the area from the

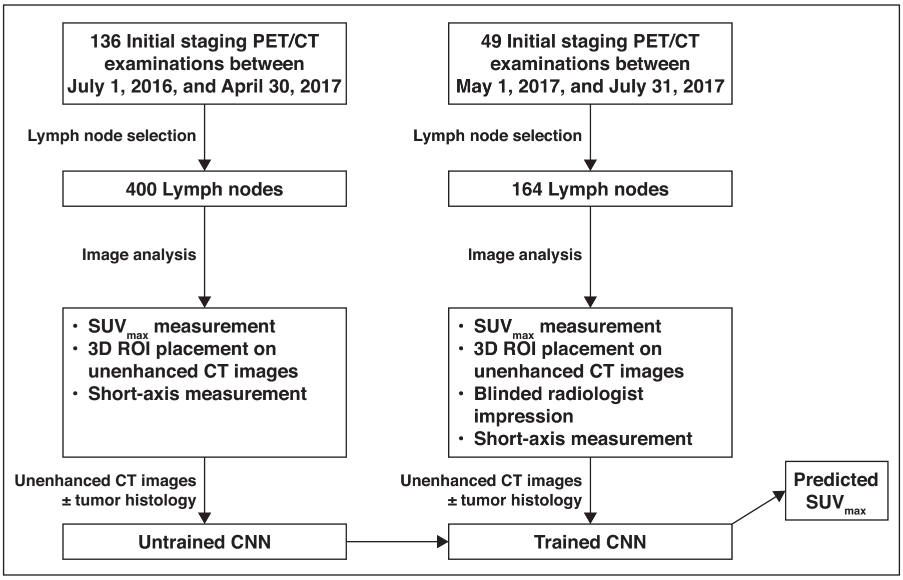
136 Initial staging PET/CT examinations between July 1, 2016, and April 30, 2017
Lymph node selection
400 Lymph nodes
Image analysis
• $SUV_{max}$ measurement
• 3D ROI placement on unenhanced CT images
• Short-axis measurement
Unenhanced CT images ± tumor histology
Untrained CNN
49 Initial staging PET/CT examinations between May 1, 2017, and July 31, 2017
Lymph node selection
164 Lymph nodes
Image analysis
• $SUV_{max}$ measurement
• 3D ROI placement on unenhanced CT images
• Blinded radiologist impression
• Short-axis measurement
Unenhanced CT images ± tumor histology
Trained CNN
Predicted $SUV_{max}$

**Fig. 1**—Flowchart of study. $SUV_{max}$ = maximum standardized uptake value, CNN = convolutional neural network.

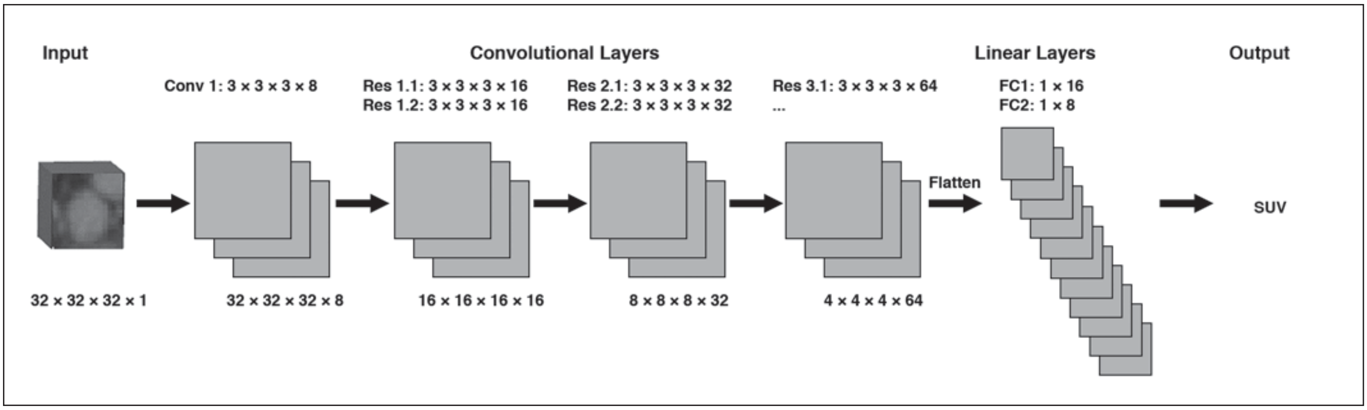

**Fig. 2**—Schematic of convolutional neural network architecture. Numbers above boxes denoting layers indicate filter types and sizes used, and numbers below boxes denoting layers indicate size of input feature maps. Each successive decrease in feature map size is attained by applying convolution operation with stride 2 as first layer in each residual operation. Conv = filter types corresponding to standard convolutional layer followed by batch normalization and rectified linear unit nonlinearity, Res = filter types corresponding to residual layers with two embedded convolutional operations followed by batch normalization and rectified linear unit nonlinearity, FC = fully connected layer, SUV = standardized uptake value.

vertex to the proximal femur were obtained while the patients were in the supine position. Fluorine-18-FDG PET/CT was performed 50 minutes after IV injection of 296–370 MBq of FDG. During the waiting period, patients rested in a dark and quiet room. The acquisition time for PET images was 4 minutes per bed position. CT images were also obtained from the patient's integrated FDG PET/CT with use of a standardized protocol and the following parameters: tube voltage, 120 kV; a variable (modulated) tube current; tube rotation time, 0.5 second per rotation; tilt, 0; and slice thickness, 4 mm. Patients were allowed to breathe normally during the procedure. Attenuation correction was performed with the unenhanced CT images.

### Image Analysis

After the lymph node database was locked, one radiologist subsequently reviewed the attenuation-corrected PET/CT fusion images in three planes (the transaxial, coronal, and sagittal planes) on a diagnostic imaging workstation (AW Workstation 3.2, GE Healthcare). The volumetric $\text{SUV}_{\text{max}}$ was recorded for each selected lymph node. Anonymized unenhanced CT DICOM images were downloaded to a password-protected external hard drive and were loaded into an open-source software platform for medical image informatics and analysis (3D Slicer, version 4.0). A spherical ROI with a radius of 2.4 cm was placed in the center of each selected lymph node and was saved. This served as the input to the preprocessing step.

### Convolutional Neural Network Architecture and Training

For the training set, data augmentation was used to improve regularization performance by exposing the network input to multiple small variations of each lymph node. The network was thus allowed to learn to marginalize random noise introduced by minor shifts in lymph node positioning as well as slight differences in acquisition parameters. Full details regarding image preprocessing are provided in Appendix S1 (which can be viewed in the *AJR* electronic supplement to this article, available at www.ajronline.org). Network inputs consisted of a bounding cube (32 × 32 × 32 mm) centered around each segmented lymph node based on an ROI centroid. A 3D CNN with 14 hidden layers and a linear regression output (predicted $\text{SUV}_{\text{max}}$) was designed using the Python programming language (version 3.5, Python Software Foundation) and TensorFlow (version 1.4, Python Software Foundation) open-source software library. The CNN was based on a novel 3D neural network architecture using multiple residual connections (Fig. 2). Run-time regularization techniques were used, including L2 regularization to limit the square magnitude of weights and dropout to limit unit coadaptation. The CNN was run with and without the input of the histopathologic diagnosis of each patient's primary malignancy. The histologic information was integrated by converting tumor histologic type, organ of origin, and a histologic modifier for further differentiating mucinous, ductal, lobular, and signet ring adenocarcinomas into three separate groups including multiple numeric categories. The categories were converted to a 1 × 8 array in which the first four entries corresponded to the tumor histologic type, the next three corresponded to the organ of origin, and the final entry in the vector corresponded to the histologic modifier. The vector was subsequently concatenated as a 1 × 8 array to the output of the first fully connected layer. Full-layer dropout was applied to the histologic input variables. Similar to weight dropout, full-layer dropout allows the network to learn robustness to missing inputs—in this case, clinical inputs. This has a twofold effect in this case, allowing the network to make predictions even in the absence of clinical information and preventing the network from overfitting on the clinical data.

The output of the network was a single number that was interpreted as the predicted $\text{SUV}_{\text{max}}$ of the network on the basis of the inputs. The Euclidian (L2) distance was used as the objective function of the network to penalize large outliers in the data. Network hyperparameters were fine-tuned on the basis of performance using a validation set composed of 20% of the training samples. The Manhattan (L1) distance was used as the primary performance metric for tuning network hyperparameters and analyzing network performance. After final hyperparameters were tuned on the validation subset of the training cohort, the network was trained on the entire training set. Additional network testing and implementation details are provided in Appendix S1.

### Testing Phase

The trained CNN was tested on the separate testing cohort of lymph nodes, and the predicted $\text{SUV}_{\text{max}}$ was recorded. Two different board-certified radiologists, both blinded to the PET images and CNN prediction, separately classified these lymph nodes as FDG avid or not FDG avid on the basis of their appearance on unenhanced CT. Short-axis measurements were recorded and were used to predict FDG avidity on the basis of a 1-cm cutoff.

### Statistical Analysis

The Wilcoxon rank sum test was used to compare the training and testing sets in terms of the median $\text{SUV}_{\text{max}}$ and the nodal short-axis measurement. An association between the real $\text{SUV}_{\text{max}}$ and the predicted $\text{SUV}_{\text{max}}$ was tested using linear re-

### TABLE 1: Patient and Lymph Node Characteristics

| Characteristic | Training | Testing | $p$ |
|---|---|---|---|
| Patient age (y) | 67 (20–93) | 67.5 (60.5–77.0) | 0.3677 |
| No. of lymph nodes per patient | 2 (1–4) | 2 (1–3) | 0.3208 |
| Lymph node $SUV_{max}$ | 6.8 (2.7–11.6) | 3.5 (1.9–8.6) | < 0.0001 |
| Lymph node short-axis measurement (cm) | 11 (9–16) | 10 (7–14) | < 0.0001 |
| Solid tumors (% of all tumors) | 71 | 87 | < 0.0001 |

Note—Except where otherwise indicated, data are median value (interquartile range). $SUV_{max}$ = maximum standardized uptake value.

gression models. For the purposes of binary classification, any lymph node with an $SUV_{max}$ of 2.5 or greater was defined as FDG avid. Interreader variability for radiologist prediction of FDG avidity was tested using the kappa coefficient. The association between nodal FDG avidity and the CNN-predicted $SUV_{max}$ was tested using a logistic regression model generating ROC curves and AUC values before and after including the nodal short axis and the radiologist's qualitative assessment as covariates. All statistical analysis was performed using statistical software (SAS, version 9.4, SAS Institute).

## Results

A total of 400 lymph nodes in 136 patients (median age, 67 years; interquartile range [IQR], 20–93 years) were used in the training phase, with a median of two lymph nodes selected per patient (IQR, one to four lymph nodes per patient). The median $SUV_{max}$ of the lymph nodes was 6.8 (IQR, 2.7–11.6), with 94 lymph nodes having an $SUV_{max}$ of less than 2.5. The median short-axis length of the lymph nodes was 1.1 cm, (IQR, 0.9–1.6 cm), with 78 lymph nodes having a short-axis length of less than 1.0 cm. The primary histologic types of the lymph nodes were as follows: adenocarcinoma ($n$ = 190), squamous cell carcinoma ($n$ = 62), diffuse large B-cell lymphoma ($n$ = 51), follicular lymphoma ($n$ = 35), Hodgkin lymphoma ($n$ = 15), small cell carcinoma ($n$ = 12), mantle cell lymphoma ($n$ = 11), melanoma ($n$ = 9), sarcoma ($n$ = 5), acute lymphoblastic leukemia ($n$ = 3), germ cell ($n$ = 2), marginal zone lymphoma ($n$ = 1), histiocytosis ($n$ = 1), and unknown type ($n$ = 3). The location of the lymph nodes was described as cervical ($n$ = 65), axillary ($n$ = 75), mediastinal or hilar ($n$ = 88), retroperitoneal ($n$ = 55), mesenteric ($n$ = 21), iliac ($n$ = 21), or inguinal ($n$ = 75).

A total of 164 lymph nodes in 49 patients (median age, 67.5 years; IQR, 60.5–77.0 years) were used in the testing phase, with a median of two lymph nodes selected per patient (IQR, one to three lymph nodes per patient). The median $SUV_{max}$ of the lymph nodes was 3.5 (IQR, 1.9–8.6), with 62 lymph nodes having an $SUV_{max}$ of less than 2.5. The median short-axis length of the lymph nodes was 1.0 cm (IQR, 0.7–1.4 cm), with 94 lymph nodes having a short-axis length of less than 1.0 cm. The primary histologic types of the lymph nodes were as follows: adenocarcinoma ($n$ = 80), squamous cell carcinoma ($n$ = 46), small cell carcinoma ($n$ = 15), diffuse large B-cell lymphoma ($n$ = 10), plasmacytoid ($n$ = 5), mantle cell lymphoma ($n$ = 4), marginal zone lymphoma ($n$ = 2), and adenoid cystic ($n$ = 2). The location of the lymph nodes was described as cervical ($n$ = 26), axillary ($n$ = 22), mediastinal or hilar ($n$ = 61), retroperitoneal ($n$ = 15), mesenteric ($n$ = 2), iliac ($n$ = 12), or inguinal ($n$ = 26).

Patient demographic characteristics and lymph node characteristics for both the training and testing groups are summarized in Table 1. Figure 3 shows the distribution of $SUV_{max}$ between the two cohorts. Details regarding organ of origin, primary histologic type, and additional histologic modifiers are presented in Figs. S2–S5 and Table S6 (which can be viewed in Appendix S1, the *AJR* electronic supplement to this article, available at www.ajronline.org).

With CT data and primary cancer histologic type used as input, the CNN-predicted $SUV_{max}$ was directly associated with the measured $SUV_{max}$ ($p < 0.0001$, β estimate = 0.83) (Fig. 4). A one-unit increase in the predicted $SUV_{max}$ was associated with an increase of 0.83 in the measured $SUV_{max}$ (model 1 in Table 2). The coefficient of determination ($R^2$) was 0.35 such that 35% of the variability of the measured $SUV_{max}$ could be attributed to the CNN-predicted $SUV_{max}$. Combining the nodal short axis measurement and the radiologist's subjective assessment in a regression model improved the $R^2$ to 0.47 and 0.52, respectively. In both of these multivariable linear regression models, the CNN-predicted

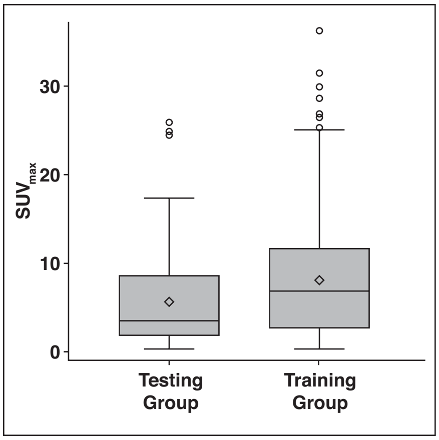

**Fig. 3**—Box-and-whisker plot of distribution of maximum standardized uptake value ($SUV_{max}$). Horizontal lines within boxes denote mean values, vertical lines and whiskers denote 95% CIs, circles denote outliers, and diamonds denote median.

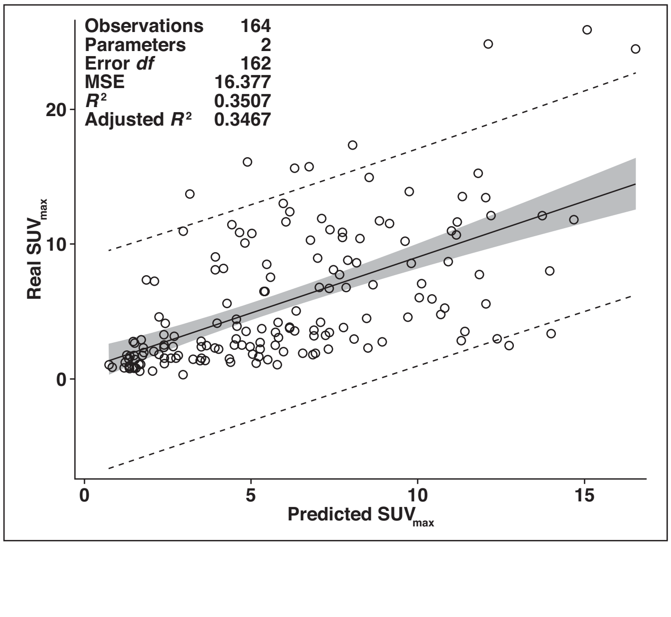

**Fig. 4**—Fit plot of association of convolutional neural network (CNN)–predicted maximum standardized uptake value ($SUV_{max}$) with real $SUV_{max}$. Solid line denotes fit, dashed line denoted 95% prediction limits, gray-shaded area denoted 95% CI, and circles denote data points. MSE = mean squared error.

**TABLE 2: Change in Real Maximum Standardized Uptake Value ($SUV_{max}$)**

| Model | Change in Real $SUV_{max}$ | 95% CI | *p* |
|---|---|---|---|
| Model 1: CNN-predicted $SUV_{max}$ | 0.83 | 0.65–1.00 | < 0.0001 |
| Model 2: model 1 and lymph node short axis | 0.46 | 0.27–0.65 | < 0.0001 |
| Model 3: model 2 and radiologist's assessment | 0.37 | 0.18–0.57 | 0.0002 |

Note—Changes in real $SUV_{max}$ resulted from one-unit increase in $SUV_{max}$ predicted by convolutional neural network (CNN) in univariable and multivariable models.

**TABLE 3: Change in Odds Ratio of Nodal FDG Avidity**

| Model | Odds Ratio | 95% CI | *p* |
|---|---|---|---|
| Model 1: CNN-predicted $SUV_{max}$ | 1.701 | 1.43–2.02 | < 0.0001 |
| Model 2: model 1 and lymph node short axis | 1.513 | 1.22–1.87 | 0.0001 |
| Model 3: model 2 and radiologist's assessment | 1.464 | 1.16–1.84 | 0.0011 |

Note—Changes in odds ratio of nodal FDG avidity resulted from one-unit increase in maximum standardized uptake value ($SUV_{max}$) predicted by convolutional neural network (CNN) in univariable and multivariable models.

$SUV_{max}$ continued to be an independent predictor of the measured $SUV_{max}$, as is shown in models 2 and 3 in Table 2.

For a binary outcome (i.e., predicting whether the measured $SUV_{max}$ was ≥ 2.5), the CNN-predicted $SUV_{max}$ was directly associated with FDG avidity ($p < 0.0001$, AUC value, 0.85); a one-unit increase in the predicted $SUV_{max}$ was associated with a 70% increase in the odds of a lymph node being FDG avid (model 1 in Table 3). The prediction of nodal FDG avidity improved after the nodal short axis was included in a multivariable logistic regression model (model 2 in Table 3; AUC value, 0.92) and after including the nodal short axis and the radiologist's prediction of FDG avidity (model 3 in Table 3; AUC value, 0.95). Figure 5 depicts the various ROCs.

The short-axis measurement of the lymph node was directly associated with the real $SUV_{max}$ in a linear regression model ($p < 0.0001$, $R^2 = 0.41$) and with nodal FDG avidity in logistic regression model (i.e., SUV ≥ 2.5; AUC value, 0.88).

There was good interreader agreement for the binary prediction of FDG avidity by the two radiologists ($\kappa = 0.68$, $p < 0.0001$). The sensitivity and specificity for reader 1 were 96.4% and 72.5%, respectively. The sensitivity and specificity for reader 2 were 88.2% and 79.0%, respectively. The diagnostic accuracy of both radiologists was identical (85%) and was similar to the performance of the CNN-predicted $SUV_{max}$ (AUC value, 0.85).

## Discussion

The results of the present study support the hypothesis that a CNN can be trained to use lymph node data from the unenhanced CT portion of PET/CT examinations of patients with cancer to predict the $SUV_{max}$. When an $SUV_{max}$ cutoff of 2.5 or greater was defined for a binary outcome, a process that radiologists perform in day-to-day clinical practice, the CNN was able to perform with high accuracy, potentially serving as a supportive diagnostic tool for the interpreting radiologist. We chose an $SUV_{max}$ cutoff of 2.5 because it was supported in a previous study, yielding a sensitivity, specificity, and negative predictive value of 89%, 84%, and 96%, respectively, in the setting of mediastinal lymph nodes in patients with lung cancer [23]. However, it is important to note that different cancers will have different FDG-uptake patterns and that no single cutoff provides the best sensitivity and specificity. The CNN output was a predictor of nodal FDG avidity independent of lymph node size. Logistic regression modeling showed incremental improvement when CNN output was combined with the short-axis measurement and the radiologist's qualitative assessment.

Classic machine learning techniques have been widely studied in radiologic imaging analysis [28, 30, 31]. Using these techniques, researchers have attempted to make diagnoses [31, 32], predict tumor grade [30, 33–35], and predict patient outcome [35, 36]. The limitation of these older machine learning techniques is the need for operators to prespecify the variables to be extracted from the images, when in fact there may be many other variables that have not yet been defined previously by human operators. A newer approach known as deep learning may use a model such as a CNN [37, 38]. CNNs are able to take in datasets, including images, and are able to classify data through a hierarchic process of convolution. This occurs without prespecifying discriminating image variables. Deep learning is increasingly being used in many industries. In radiologic image analysis alone, recent studies have shown that deep learning has great potential [18, 39, 40]. Two major limitations of deep learning are the need for large datasets to prevent overfitting and the nontransparency associated with feature selection by the CNN. The first limitation may be overcome through data augmentation in which data are manipu-

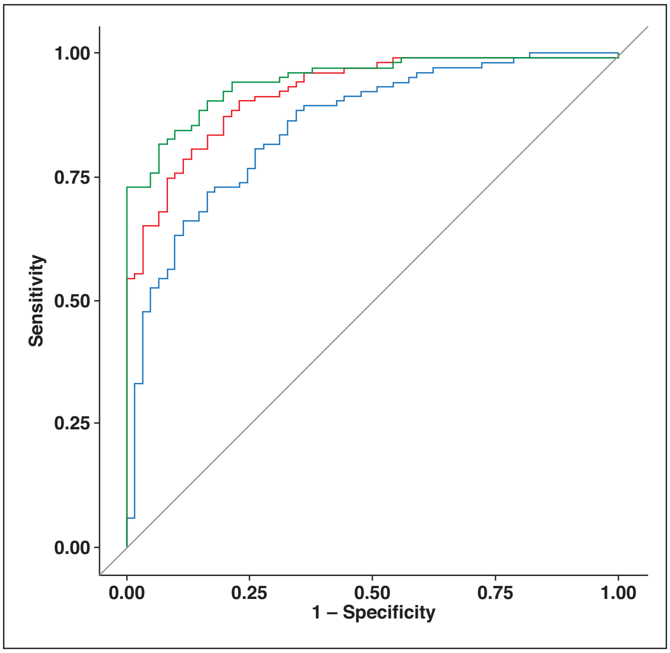

**Fig. 5**—Line graph of ROCs for predicting FDG avidity (maximum standardized uptake value, ≥ 2.5). Blue line denotes convolutional neural network (CNN) with image and cancer histologic input; red line represents CNN with image and cancer histologic input and short-axis measurement greater than 1 cm; and green line denotes CNN with image and cancer histologic input, short-axis measurement greater than 1 cm, and radiologist's qualitative assessment. Diagonal line indicates 0.50.

lated to artificially enlarge the dataset. Several methods have been developed to deal with the latter limitation, including utilization of guided back propagation, which displays the pixels on the input image that the network gives the most weight to when making a final decision; visualization of layer representations in hidden layers; and class activation mapping techniques that highlight ROIs in the input pictures that contribute to positive identifications of a class [41–44].

In the present study, we intentionally used the same-size spherical ROI for all lymph nodes rather than manual segmentation of each lymph node. In this way, we believed that we could minimize the effect of lymph node size in the model and focus the CNN on lymph node texture and shape. The improved performance of short-axis dimension and the CNN in our analysis suggests that the CNN did not use size as a dominant feature. However, size is a major criteria for discriminating benign from malignant lymph nodes [26]. It is interesting to hypothesize whether manual segmentation of each lymph node, rather than use of a predetermined spherical ROI, would lead to an improved outcome, presumably as a result of incorporation of lymph node size into the CNN classification algorithm.

Bayanti et al. [17] studied 72 mediastinal lymph nodes with pathologic correlation in 43 patients with lung cancer. They used classic machine learning (the support vector machine model) and prespecified texture characteristics extracted from unenhanced CT images. They reported a sensitivity of 81% and a specificity of 80% (AUC value, 0.87; $p < 0.0001$). Of importance, the study design used internal cross-validation (which is prone to the limitation of overfitting) rather than a separate testing cohort. Furthermore, the requirement for pathologic diagnosis introduced a selection bias. Wang et al. [18] examined 1397 lymph nodes in 168 patients with non–small cell lung cancer and attempted to classify lymph nodes as benign or malignant with the use of both unenhanced CT and PET images and surgical pathologic findings as the reference standard. Wang and colleagues compared the performance of classic machine learning, radiologist qualitative assessment, and deep learning with use of a CNN. Their results show similar performance of all three methods, with AUC values of 0.87–0.92. Our study differed in several key ways. Wang and colleagues used three axial slices from both the unenhanced CT and PET images, whereas we relied on 3D sampling of the lymph nodes on unenhanced CT only and attempted to predict the $SUV_{max}$. Of importance, we trained and validated our CNN using one dataset and tested it using a separate dataset, one not previously seen by the CNN, rather than using cross-validation.

The present study had several key limitations. First, the reference standard $SUV_{max}$ is a common surrogate for malignancy but can also be elevated in benign conditions. Therefore, although predicting the $SUV_{max}$ may be clinically meaningful, it does not equate to pathologic confirmation of malignancy. Second, there is no single $SUV_{max}$ cutoff that can be used accurately for all malignancies. Third, although selection of lymph nodes was done without knowledge of the primary malignancy and without viewing the PET images, a selection bias may have existed. In this regard, our sole criteria for lymph node selection was that the lymph node was not obscured by adjacent soft-tissue structures. Fourth, we chose to analyze the unenhanced CT images acquired during PET/CT examinations as opposed to the contrast-enhanced CT images obtained at a different point in time but close to the time of the PET/CT examination. This ensured that any potential treatment effects or tumor growth could not develop between our reference standard (PET) and the testing data (CT). Theoretically, contrast-enhanced CT may allow different and potentially more textural information to be extracted, thus improving the CNN performance. This should be explored in future studies.

## Conclusion

In patients with cancer, deep learning using a CNN is able to predict $SUV_{max}$ with moderate accuracy on the basis of unenhanced CT images and cancer histologic subtype. Short-axis measurement and radiologist qualitative assessment are complementary, and when combined into a predictive model, perform with a high accuracy in classifying lymph nodes as FDG avid versus not FDG avid (cutoff value, 2.5). Using this CNN algorithm for routine unenhanced CT studies may provide additional information to the interpreting radiologist when PET/CT is not immediately available. More importantly, future research may focus on combining machine learning analysis of lymph nodes in tandem with PET SUV data into a single predictive model for malignancy. As mentioned in the introduction to this article, this will necessitate obtaining robust, unbiased pathologic data on patients' lymph nodes.

## References

1. Amin MB, Edge SB, Greene FL, et al., eds. *AJCC cancer staging manual*, 8th ed. Chicago, IL: American College of Surgeons, 2018
2. National Comprehensive Cancer Network (NCCN). Cancer staging guide. NCCN website. www.nccn.org/patients/resources/diagnosis/staging.aspx. Accessed October 31, 2018
3. Koh DM, Cook GJ, Husband JE. New horizons in oncologic imaging. *N Engl J Med* 2003; 348:2487–2488
4. Johnson SA, Kumar A, Matasar MJ, Schoder H, Rademaker J. Imaging for staging and response assessment in lymphoma. *Radiology* 2015; 276:323–338
5. Heacock L, Weissbrot J, Raad R, et al. PET/MRI for the evaluation of patients with lymphoma: initial observations. *AJR* 2015; 204:842–848
6. Sun J, Li B, Li CJ, et al. Computed tomography versus magnetic resonance imaging for diagnosing cervical lymph node metastasis of head and neck cancer: a systematic review and meta-analysis. *OncoTargets Ther* 2015; 8:1291–1313
7. Choi HJ, Roh JW, Seo SS, et al. Comparison of the accuracy of magnetic resonance imaging and positron emission tomography/computed tomography in the presurgical detection of lymph node metastases in patients with uterine cervical carcinoma: a prospective study. *Cancer* 2006; 106:914–922
8. Scaranelo AM, Eiada R, Jacks LM, Kulkarni SR, Crystal P. Accuracy of unenhanced MR imaging in the detection of axillary lymph node metastasis: study of reproducibility and reliability. *Radiology* 2012; 262:425–434
9. Hövels AM, Heesakkers RA, Adang EM, et al. The diagnostic accuracy of CT and MRI in the staging of pelvic lymph nodes in patients with prostate cancer: a meta-analysis. *Clin Radiol* 2008; 63:387–395
10. Wahl RL, Siegel BA, Coleman RE, Gatsonis CG; PET Study Group. Prospective multicenter study of axillary nodal staging by positron emission tomography in breast cancer: a report of the staging breast cancer with PET Study Group. *J Clin Oncol* 2004; 22:277–285
11. Guller U, Nitzsche E, Moch H, Zuber M. Is positron emission tomography an accurate non-invasive alternative to sentinel lymph node biopsy in breast cancer patients? *J Natl Cancer Inst* 2003; 95:1040–1043
12. Kitajima K, Murakami K, Yamasaki E, et al. Accuracy of $^{18}$F-FDG PET/CT in detecting pelvic and paraaortic lymph node metastasis in patients with endometrial cancer. *AJR* 2008; 190:1652–1658
13. Kwak JY, Kim JS, Kim HJ, Ha HK, Yu CS, Kim JC. Diagnostic value of FDG-PET/CT for lymph node metastasis of colorectal cancer. *World J Surg* 2012; 36:1898–1905

14. Betancourt Cuellar SL, Sabloff B, Carter BW, et al. Early clinical esophageal adenocarcinoma (cT1): utility of CT in regional nodal metastasis detection and can the clinical accuracy be improved? *Eur J Radiol* 2017; 88:56–60
15. Billè A, Okiror L, Skanjeti A, et al. Evaluation of integrated positron emission tomography and computed tomography accuracy in detecting lymph node metastasis in patients with adenocarcinoma vs squamous cell carcinoma. *Eur J Cardiothorac Surg* 2013; 43:574–579
16. Reinhardt MJ, Ehritt-Braun C, Vogelgesang D, et al. Metastatic lymph nodes in patients with cervical cancer: detection with MR imaging and FDG PET. *Radiology* 2001; 218:776–782
17. Bayanati H, Thornhill RE, Souza CA, et al. Quantitative CT texture and shape analysis: can it differentiate benign and malignant mediastinal lymph nodes in patients with primary lung cancer? *Eur Radiol* 2015; 25:480–487
18. Wang H, Zhou Z, Li Y, et al. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from $^{18}$F-FDG PET/CT images. *EJNMMI Res* 2017; 7:11
19. Li M, Wu N, Liu Y, et al. Regional nodal staging with $^{18}$F-FDG PET-CT in non-small cell lung cancer: additional diagnostic value of CT attenuation and dual-time-point imaging. *Eur J Radiol* 2012; 81:1886–1890
20. Flechsig P, Frank P, Kratochwil C, et al. Radiomic analysis using density threshold for FDG-PET/CT-based N-staging in lung cancer patients. *Mol Imaging Biol* 2017; 19:315–322
21. Giesel FL, Schneider F, Kratochwil C, et al. Correlation between $SUV_{max}$ and CT radiomic analysis using lymph node density in PET/CT-based lymph node staging. *J Nucl Med* 2017; 58:282–287
22. Kitajima K, Fukushima K, Miyoshi Y, et al. Diagnostic and prognostic value of $^{18}$F-FDG PET/CT for axillary lymph node staging in patients with breast cancer. *Jpn J Radiol* 2016; 34:220–228
23. Hellwig D, Graeter TP, Ukena D, et al. 18F-FDG PET for mediastinal staging of lung cancer: which SUV threshold makes sense? *J Nucl Med* 2007; 48:1761–1766
24. Lv YL, Yuan DM, Wang K, et al. Diagnostic performance of integrated positron emission tomography/computed tomography for mediastinal lymph node staging in non-small cell lung cancer: a bivariate systematic review and meta-analysis. *J Thorac Oncol* 2011; 6:1350–1358
25. Toloza EM, Harpole L, McCrory DC. Noninvasive staging of non-small cell lung cancer: a review of the current evidence. *Chest* 2003; 123:137S–146S
26. Schwartz LH, Bogaerts J, Ford R, et al. Evaluation of lymph nodes with RECIST 1.1. *Eur J Cancer* 2009; 45:261–267
27. Gao X, Chu C, Li Y, et al. The method and efficacy of support vector machine classifiers based on texture features and multi-resolution histogram from $^{18}$F-FDG PET-CT images for the evaluation of mediastinal lymph nodes in patients with lung cancer. *Eur J Radiol* 2015; 84:312–317
28. Giger ML. Machine learning in medical imaging. *J Am Coll Radiol* 2018; 15:512–520
29. Murphy JD, Chisholm KM, Daly ME, et al. Correlation between metabolic tumor volume and pathologic tumor volume in squamous cell carcinoma of the oral cavity. *Radiother Oncol* 2011; 101:356–361
30. Yasaka K, Akai H, Abe O, Ohtomo K, Kiryu S. Quantitative computed tomography texture analyses for anterior mediastinal masses: differentiation between solid masses and cysts. *Eur J Radiol* 2018; 100:85–91
31. Cohen JG, Reymond E, Medici M, et al. CT-texture analysis of subsolid nodules for differentiating invasive from in-situ and minimally invasive lung adenocarcinoma subtypes. *Diagn Interv Imaging* 2018; 99:291–299
32. Lee-Felker SA, Felker ER, Tan N, et al. Qualitative and quantitative MDCT features for differentiating clear cell renal cell carcinoma from other solid renal cortical masses. *AJR* 2014; 203:[web]W516–W524
33. Algohary A, Viswanath S, Shiradkar R, et al. Radiomic features on MRI enable risk categorization of prostate cancer patients on active surveillance: preliminary findings. *J Magn Reson Imaging* 2018 Feb 22 [Epub ahead of print]
34. Kierans AS, Rusinek H, Lee A, et al. Textural differences in apparent diffusion coefficient between low- and high-stage clear cell renal cell carcinoma. *AJR* 2014; 203:[web]W637–W644
35. Lubner MG, Stabo N, Abel EJ, Del Rio AM, Pickhardt PJ. CT textural analysis of large primary renal cell carcinomas: pretreatment tumor heterogeneity correlates with histologic findings and clinical outcomes. *AJR* 2016; 207:96–105
36. Attiyeh MA, Chakraborty J, Doussot A, et al. Survival prediction in pancreatic ductal adenocarcinoma by quantitative computed tomography image analysis. *Ann Surg Oncol* 2018; 25:1034–1042
37. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; 521:436–444
38. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM* 2017; 60:84–90
39. Yasaka K, Akai H, Abe O, Kiryu S. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. *Radiology* 2018; 286:887–896
40. Song Q, Zhao L, Luo X, Dou X. Using deep learning for classification of lung nodules on computed tomography images. *J Healthc Eng* 2017; 2017:8314740
41. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. arXiv website. arxiv.org/pdf/1311.2901.pdf. Revised November 12, 2013. Accessed October 5, 2018
42. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv website. arxiv.org/abs/1312.6034. Revised April 19, 2014. Accessed October 5, 2018
43. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. arXiv website. arxiv.org/abs/1610.02391. Revised March 21, 2016. Accessed October 5, 2018
44. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: the all convolutional net. arXiv website. arxiv.org/abs/1412.6806. Revised April 13, 2015. Accessed October 5, 2018

**FOR YOUR INFORMATION**

A data supplement for this article can be viewed in the online version of the article at: www.ajronline.org.